Exploring Wikipedia Edits Graph

Maria Mitkina , Chase Gottlich, Kishore Vasan

Project Abstract

Mining large graphs reveals information; temporal network of the same reveal evolution. However, performing novel algorithms on these large graphs can be computationally expensive. We need methods that can provide an un-biased sample that would be representative of the underlying large network. In this work, we evaluate different random walks by crawling a large online editing network – Wikipedia.

Project Outline

Wiki-talk temporal network:

The temporal network represents Wikipedia users editing each other's Talk page. Analyzing this network reveals dynamics of user edits in a large online network. A directed edge (u, v, t) means that user u edited user v's talk page at time t. [1]

T	<V>	<E>T	<E>	<k>	<C>
2320d	1.14M	7.83M	3.3 M	5.491812	0.0014

Example of a user page and an edit

Research Questions:

1) *How does the temporal nature of the Wikipedia network affect the network statistics?*
2) *How does modifications of random walk affect the sampling performance of the entire network?*

Main Findings:

- **Clustering of the graph associated with high growth in the platform.**
- **Simple Random Walk is ineffective when sampling graphs with high tailed distribution.**
- **Re-Weighted Random Walk outperforms other methods for graph sampling.**

Trends in Summary Statistics

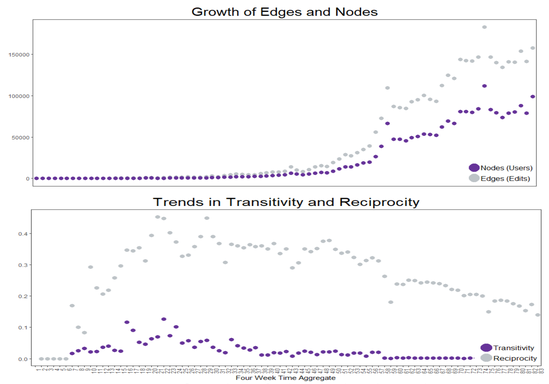

FIG 1: Summary statistics of number of edges and nodes and reciprocity and transitivity over 4-week time aggregates.

Crawling Wikipedia

- *Random Walks* (RW) allow for node re-visiting and have properties of effectively sampling a large graph. In the past this has been used in Web, P2P networks, Online Social Networks etc. [2]
- We perform different modifications of *RW* on our network and evaluate the effectiveness of sampling.
- We compare the results across three metrics of interest - node degree, number of edits, and time span of edits.
- Since the network distribution is left-skewed, we remove ~10% of low degree nodes and run the same methods.
- To measure convergence we use a ”burn in” period and use the Geweke diagnostic of MCMC: $z = \frac{E(X_a) - E(X_b)}{\sqrt{Var(X_a) + Var(X_b)}}$

Methods Description:

➢ **Simple Random Walk (RW):**
The next node is chosen uniformly at random among the current node's neighbors. $P^{RW}_{v,w} = \begin{cases} \frac{1}{k_v} & \text{if } w \text{ is a neighbor of } v, \\ 0 & \text{otherwise.} \end{cases}$

➢ **Re-Weighted Random Walk (RWRW):**
After conducting a random walk, we correct the bias of the estimator by reweighting in the end. $\hat{p}(A_i) = \frac{\sum_{u \in A_i} 1/k_u}{\sum_{u \in V} 1/k_u}$

➢ **Metropolitan Hastings Random Walk (MHRW)** :
Instead of correcting the bias after the walk, one can appropriately modify the transition probabilities so that it converges. $P^{MH}_{v,w} = \begin{cases} \frac{1}{k_v} \cdot \min(1, \frac{k_v}{k_w}) & \text{if } w \text{ is a neighbor of } v, \\ 1 - \sum_{y \neq v} P^{MH}_{v,y} & \text{if } w = v, \\ 0 & \text{otherwise.} \end{cases}$

Development of Network

Studying evolution of network reveals network dynamics. As Wikipedia nodes and edges expanded exponentially, transitivity and reciprocity metrics decreased, *signaling clustering of users,* perhaps by discipline or topic. This trend is consistent with the 'small-world phenomena' that posits large networks often consist of multiple clusters linked together by influential nodes. [3] The development of the network also follows research that characterizes typical online diffusion processes as driven by a small subset of nodes and shallow cascades leading to clustered networks as opposed to diffuse graphs. [4] The generalizability of these trends in online diffusion and whether they serve as proxies for robust platforms should be considered for future research.

Visualizing the Network

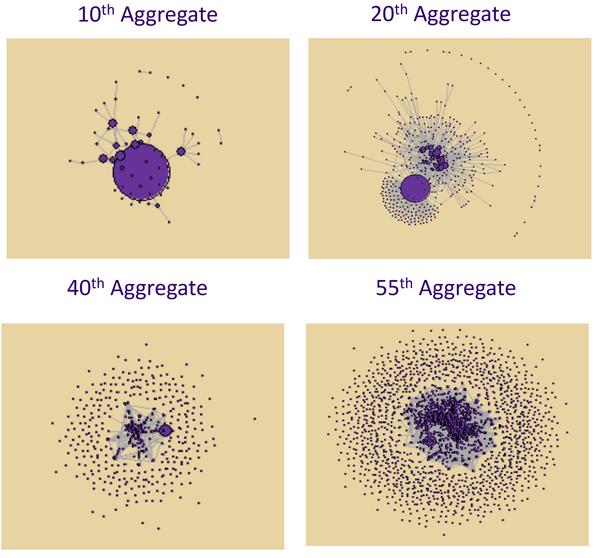

FIG 2: Visualizing the network by different time aggregates. Aggregates were over 4-week intervals from the first edit. We observe that the later we take the aggregate the stronger the evidence of super-users and clustering.

Results - Crawling

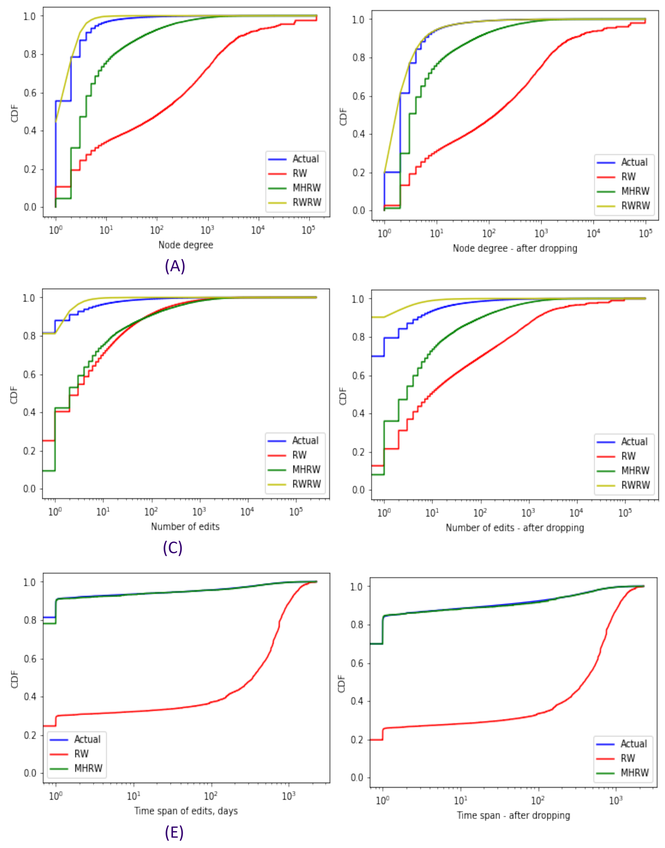

FIG 3: Performance evaluation of the graph sampling techniques using different metrics of interest. 1) RWRW tends to outperform other algorithms in estimating the properties of the network. 2) Simple RW is very bad at estimating the entire network since it oversamples higher degree nodes. 3) After removing ~10% of the nodes with low degree, MHRW is able to estimate the network better than earlier.

References:

1. Ashwin Paranjape, Austin R. Benson, and Jure Leskovec. "Motifs in Temporal Networks." *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017.
2. Gjoka, Minas, et al. "Walking in facebook: A case study of unbiased sampling of osns." *2010 Proceedings IEEE Infocom*. Ieee, 2010.
3. Milgram, S. "The small world problem". *Psychol. Today* **2**, 60–67 (1967).
4. Goel, Sharad, Duncan J. Watts, and Daniel G. Goldstein. "The Structure of Online Diffusion Networks." In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 623–638. 2012.